Statistica – Un Esempio

Un'indagine sul peso, su un campione di $n = 100$ studenti, ha prodotto il seguente risultato. I pesi p sono espressi in Kg e sono stati raggruppati in cinque *classi* di peso.

classe (peso in Kg)	f_i	f_i/n
$60 \leq p \leq 62$	5	0.05
$63 \leq p \leq 65$	18	0.18
$66 \leq p \leq 68$	42	0.42
$69 \leq p \leq 71$	27	0.27
$72 \leq p \leq 74$	8	0.08
	100	1.00

Sono riportate le frequenze assolute f_i (numero di individui appartenenti alla classe di peso i-sima) e le frequenze relative f_i/n.

Le classi sono di uguale ampiezza, ma non sono contigue

Statistica – Un Esempio

Estendiamo i confini di ciascuna classe in modo simmetrico di 0.5 Kg. La popolazione non è cambiata e possiamo rappresentare i dati, in classi contigue, come segue:

classe (peso in Kg)	r_i	f_i	f_i/n
$59.5 \leq p < 62.5$	61	5	0.05
$62.5 \leq p < 65.5$	64	18	0.18
$65.5 \leq p < 68.5$	67	42	0.42
$68.8 \leq p < 71.5$	70	27	0.27
$71.5 \leq p < 74.5$	73	8	0.08
		100	1.00

Supponendo che gli individui di una classe siano distribuiti *uniformemente* al suo interno, è naturale associare a ciascuna classe, come *rappresentante*, il valore centrale r_i della classe stessa.

Calcolo della Media

Come si può calcolare la media dei dati conoscendo solo un'informazione parziale *(per classi)* sulle frequenze?

Occorre formulare un'ipotesi su come i dati si distribuiscono all'*interno* di ogni classe. In assenza di ulteriori informazioni, è ragionevole congetturare che gli elementi appartenenti ad una classe si distribuiscano *uniformemente* al suo interno.

È naturale associare ad ogni classe un *rappresentante*: il valore centrale della classe.

r_i	61	64	67	70	73
f_i	5	18	42	27	8

Ai fini del calcolo della media si utilizzano solo i rappresentanti r_i:

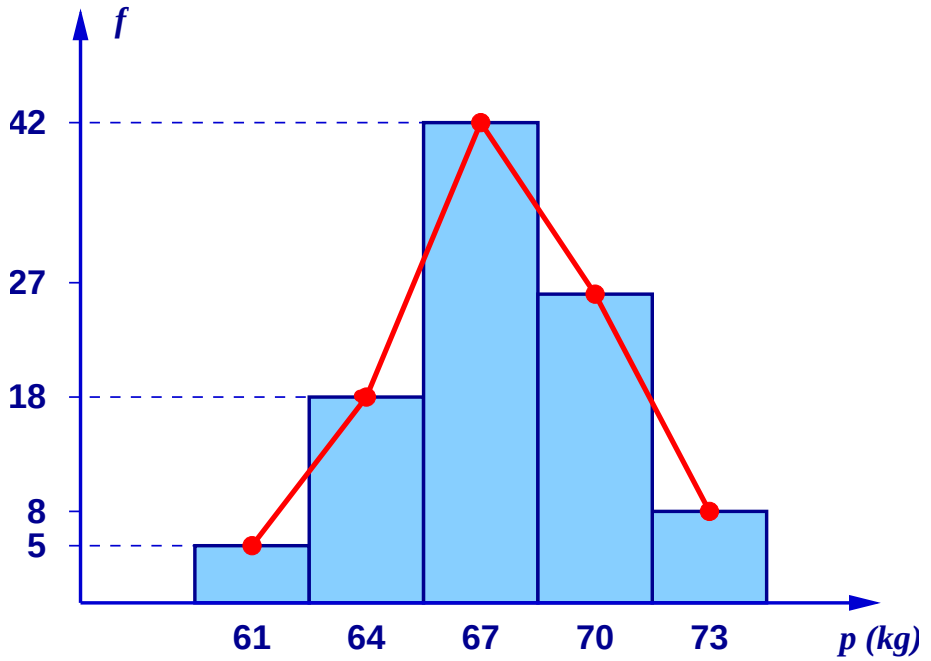
$$\overline{p} = \frac{5 \cdot 61 + 18 \cdot 64 + 42 \cdot 67 + 27 \cdot 70 + 8 \cdot 73}{100} = 67.45\,\mathrm{Kg}$$

Poligono di Frequenza

Distribuzione delle frequenze

classe	r_i	f_i
$59.5 \leq p < 62.5$	61	5
$62.5 \leq p < 65.5$	64	18
$65.5 \leq p < 68.5$	67	42
$68.8 \leq p < 71.5$	70	27
$71.5 \leq p < 74.5$	73	8

- possiamo rappresentare in modo efficace le frequenze delle classi del campione mediante un istogramma (dove le aree dei rettangoli sono proporzionali alle frequenze della classe)
- unendo i punti medi • dei lati superiori dei rettangoli, si ottiene il cosiddetto poligono di frequenza



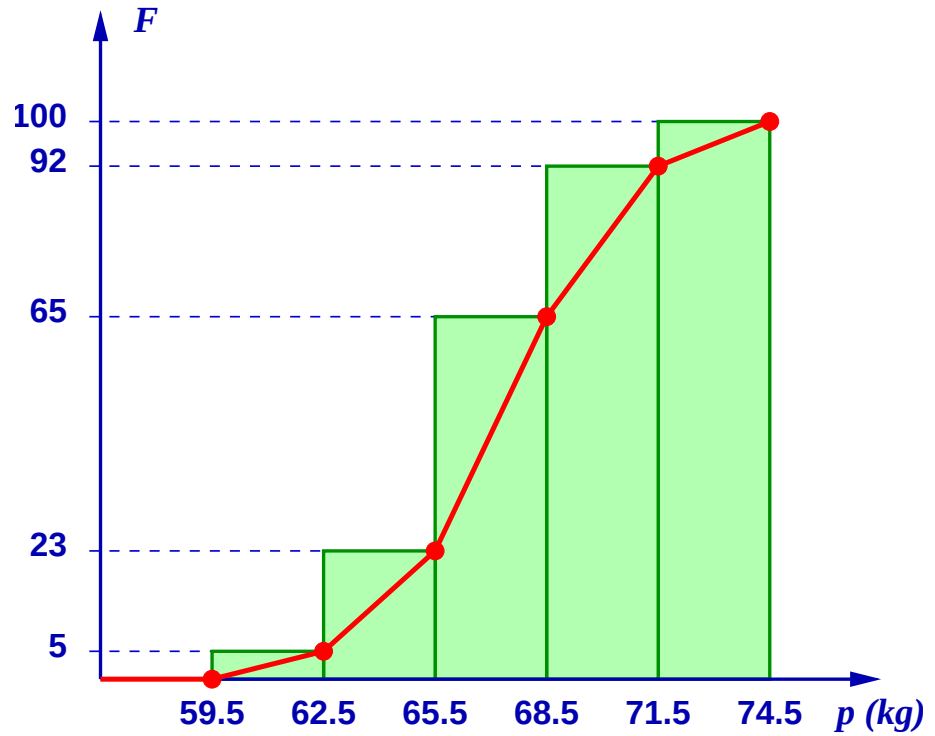
Ipotesi: classi equispaziate – distribuzione uniforme all'interno della classe

Ogiva di Frequenza

Distribuzione delle frequenze

classe	r_i	f_i	F_i
$p < 59.5$	-	0	0
$59.5 \leq p < 62.5$	61	5	5
$62.5 \leq p < 65.5$	64	18	23
$65.5 \leq p < 68.5$	67	42	65
$68.8 \leq p < 71.5$	70	27	92
$71.5 \leq p < 74.5$	73	8	100

- calcoliamo le *frequenze cumulate* F_i (F_i rappresenta il numero dei dati, che sono minori del secondo estremo della i-sima classe)
- costruiamo il diagramma cumulativo delle frequenze
- unendo i punti •, si ottiene la cosiddetta ogiva di frequenza



Calcolo della Mediana: Primo Metodo

Calcolo della mediana M_e

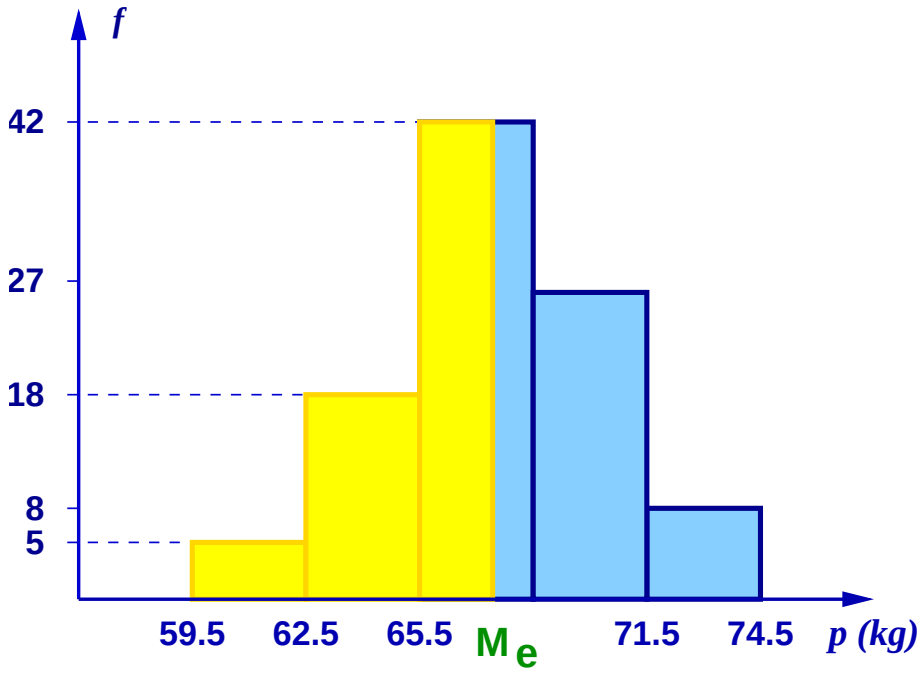
Trovare il punto M_e tale che l'area in giallo sia il 50% dell'area totale sottesa dall'istogramma delle frequenze

area totale istogramma $= 300$

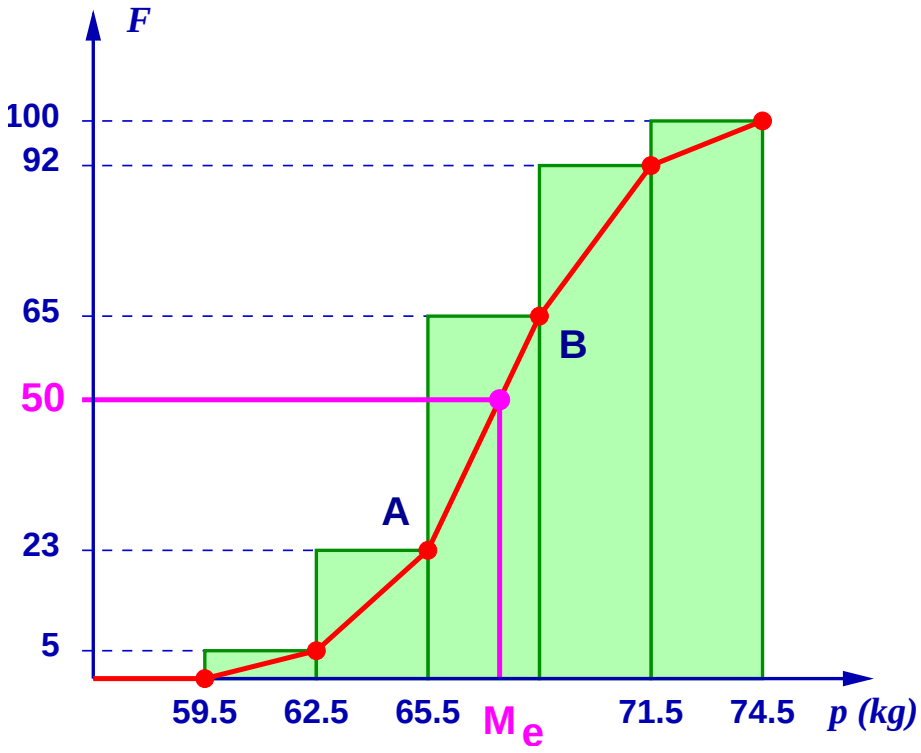
$$3 \cdot 5 + 3 \cdot 18 + (M_e - 65.5) \cdot 42 = 150$$

$$\Rightarrow \quad M_e = \frac{81}{42} + 65.5 \simeq 67.43\,\text{Kg}$$

NOTA: ricordiamo che le aree sono proporzionali alle frequenze.



Calcolo della Mediana: Secondo Metodo



Calcolo della mediana M_e

Trovare il punto di intersezione della retta $F = 50$ con l'ogiva di frequenza.

Significa trovare l'intersezione con la retta passante per i punti:

$A = (65.5\,,\,23)$ e $B = (68.5\,,\,65)$

$$\begin{cases} F = 50 \\ F = 23 + \dfrac{42}{3} \cdot (p - 65.5) \end{cases}$$

$$\Rightarrow \quad p = 65.5 + \frac{3}{42} \cdot 27 \simeq 67.43$$

Indici di Dispersione

Si cercano indici di dispersione che:

- utilizzino tutti i dati $\{x_1, x_2, \ldots, x_n\}$
- siano basati sulla nozione di scarto (distanza) dei dati
  - rispetto a un centro $d_i = |x_i - C|$
    ad esempio, rispetto alla media aritmetica $d_i = |x_i - \bar{x}|$
  - rispetto a un dato $d_i = |x_i - x_j|$

con alcune proprietà generali:

- l'indice di dispersione non deve mai essere negativo
- assume il valore 0 se i dati sono tutti uguali
- non cambia se si aggiunge una costante ai dati

Varianza

Varianza: è la media aritmetica *(semplice o ponderata)* dei quadrati degli scarti.

- Dato l'insieme di valori $\{x_1\,,\,x_2\,,\,\ldots\,,\,x_n\}$

$$\text{Var} = s^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Dato l'insieme di valori $\{x_1\,,\,x_2\,,\,\ldots\,,\,x_m\}$ con le rispettive frequenze assolute $\{f_1\,,\,f_2\,,\,\ldots\,,\,f_m\}$

$$\text{Var} = s^2 = \frac{1}{n} \cdot \sum_{i=1}^{m} f_i \cdot (x_i - \bar{x})^2 \qquad \text{dove } n = \sum_{i=1}^{m} f_i$$

Deviazione Standard

Deviazione standard *(o scarto quadratico medio)*: è la radice quadrata della varianza.

$$s = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{oppure} \quad s = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{m} f_i \cdot (x_i - \bar{x})^2}$$

Consente di avere un indice di dispersione espresso nella stessa unità di misura dei dati.

Nota: applicando una trasformazione lineare ai dati

$$y_i = ax_i + b \qquad \Rightarrow \qquad s_y^2 = a^2 s_x^2, \qquad s_y = |a|\, s_x$$

Statistiche Campionarie

Spesso gli *indici statistici* vengono applicati non all'intera *popolazione*, ma a un suo campione. Si cerca di stimare (inferenza) nel miglior modo possibile le caratteristiche dell'intera popolazione a partire dalle informazioni desunte da un campione rappresentativo.
In questo caso si utilizzano le seguenti formule modificate:

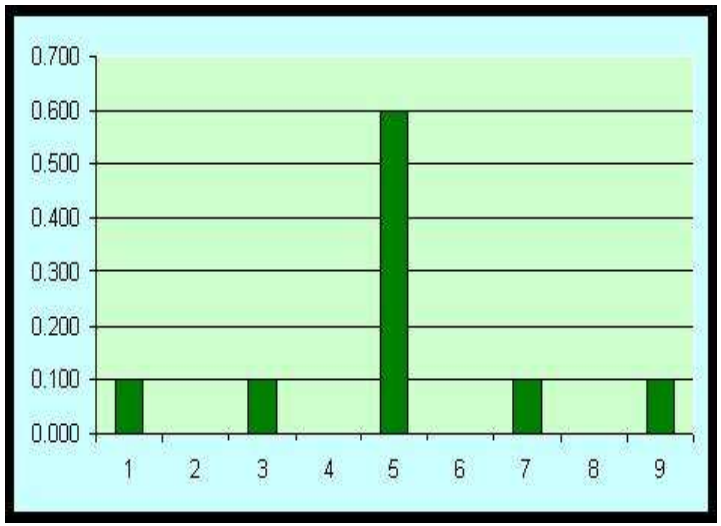
Varianza campionaria (*stimata*):

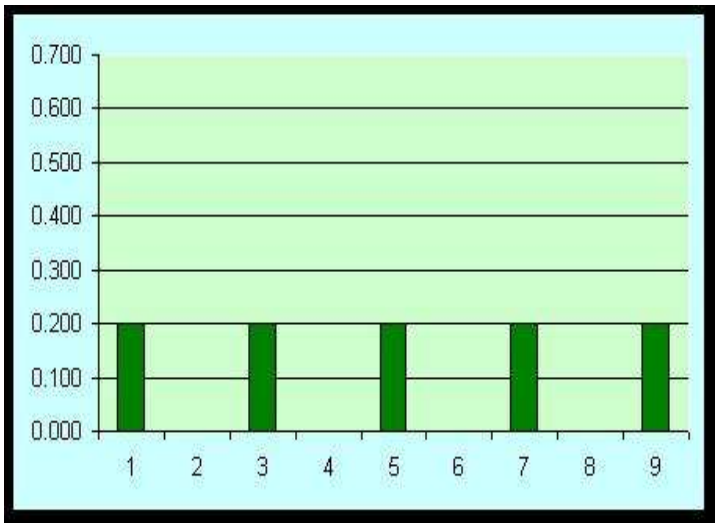
$$s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Deviazione standard campionaria (*stimata*):

$$s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Esempio Riassuntivo





Caso A

x_i	f_i	f_i/n
1	1	0.100
3	1	0.100
5	6	0.600
7	1	0.100
9	1	0.100
	10	1.000

Caso B

x_i	f_i	f_i/n
1	2	0.200
3	2	0.200
5	2	0.200
7	2	0.200
9	2	0.200
	10	1.000

Esempio Riassuntivo

Caso A

x_i	f_i	f_i/n
1	1	0.100
3	1	0.100
5	6	0.600
7	1	0.100
9	1	0.100
	10	1.000

media	5.00
mediana	5.00
varianza	4.00
varianza stimata	4.44
deviazione standard	2.00
deviazione standard stimata	2.11

Caso B

x_i	f_i	f_i/n
1	2	0.200
3	2	0.200
5	2	0.200
7	2	0.200
9	2	0.200
	10	1.000

media	5.00
mediana	5.00
varianza	8.00
varianza stimata	8.89
deviazione standard	2.83
deviazione standard stimata	2.98

Esercizi

Esercizio 1. Si consideri la seguente tabella relativa alle frequenze dei pesi in Kg di 100 individui adulti.

Peso p in Kg	f_{ass}
$50 \leq p < 55$	20
$55 \leq p < 60$	15
$60 \leq p < 65$	18
$65 \leq p < 70$	22
$70 \leq p < 75$	18
$75 \leq p < 80$	7

- le classi sono di uguale ampiezza
- supponiamo che i dati siano uniformemente distribuiti all'interno di ogni classe
- possiamo definire per ogni classe un rappresentante r_i *(class mark)*

Calcolare il peso medio e lo scarto quadratico medio.

Esercizi

Soluzione: calcoliamo la media e lo scarto quadratico medio utilizzando i valori dei rappresentanti.

Peso p in Kg	f_i	F_i	r_i
$50 \leq p < 55$	20	20	52.5
$55 \leq p < 60$	15	35	57.5
$60 \leq p < 65$	18	53	62.5
$65 \leq p < 70$	22	75	67.5
$70 \leq p < 75$	18	93	72.5
$75 \leq p < 80$	7	100	77.5

Calcoliamo il peso medio:

$$\bar{p} = \frac{1}{100}(20 \cdot 52.5 + 15 \cdot 57.5 + 18 \cdot 62.5 + 22 \cdot 67.5 + \\ + 18 \cdot 72.5 + 7 \cdot 77.5) = 63.7\,\text{Kg}$$

Esercizi

Calcoliamo la varianza e lo scarto quadratico medio:

r_i	$r_i - \bar{p}$	$(r_i - \bar{p})^2$	f_i
52.5	-11.2	125.44	20
57.5	-6.2	38.44	15
62.5	-1.2	1.44	18
67.5	3.8	14.44	22
72.5	8.8	77.44	18
77.5	13.8	190.44	7

$$s^2 = \frac{1}{100}(20 \cdot 125.44 + 15 \cdot 38.44 + 18 \cdot 1.44 + 22 \cdot 14.44 + \\ +18 \cdot 77.44 + 7 \cdot 190.44) \simeq 61.56\,\text{Kg}^2$$

$$s \simeq 7.85\,\text{Kg}$$

Media – Varianza – Deviazione Standard

$\bar{x}$ media	$\frac{1}{n} \cdot \sum_{i=1}^{n} x_i$	$\frac{1}{n} \cdot \sum_{i=1}^{m} f_i \, x_i$
s^2 varianza	$\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$	$\frac{1}{n} \cdot \sum_{i=1}^{m} f_i \cdot (x_i - \bar{x})^2$
s dev. standard	$\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}$	$\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{m} f_i \cdot (x_i - \bar{x})^2}$
s^2 campionaria	$\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$	$\frac{1}{n-1} \cdot \sum_{i=1}^{m} f_i \cdot (x_i - \bar{x})^2$
s campionaria	$\sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}$	$\sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{m} f_i \cdot (x_i - \bar{x})^2}$

Varianza – Deviazione Standard

Le espressioni della varianza (e della deviazione standard) possono essere riscritte come segue:

$$s^2 = \frac{1}{n} \cdot \left(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2 \right) \qquad \text{o} \qquad s^2 = \frac{1}{n} \cdot \left(\sum_{i=1}^{m} f_i\, x_i^2 - n\,\bar{x}^2 \right)$$

Infatti,

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i^2 - 2x_i\,\bar{x} + \bar{x}^2) = \sum_{i=1}^{n} x_i^2 - 2\bar{x} \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2 =$$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}(n\,\bar{x}) + n\,\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2$$

Esercizi

Esercizio 1. Nel rilevare l'altezza in cm di un gruppo di reclute si è ottenuta la seguente tabella delle frequenze. Calcolare media, mediana e quartili.

Soluzione:

cm	f_{ass}	f_{cum}
166	1	1
168	3	4
169	6	10
170	11	21
171	8	29
172	6	35
173	4	39
174	3	42
175	1	43
178	1	44

$n = 44$ dimensione del campione

$\bar{x} \simeq 170.9$ media

$M_e = \dfrac{x_{22} + x_{23}}{2} = 171$ mediana

$q_1 = \dfrac{x_{11} + x_{12}}{2} = 170$ primo quartile

$q_3 = \dfrac{x_{33} + x_{34}}{2} = 172$ terzo quartile

$q_3 - q_1 = 2$ distanza interquartile

La distanza interquartile è un altro indice di dispersione, legato alla nozione di mediana. La mediana suddivide l'insieme dei dati ordinati $\{x_i\}$ in due parti ugualmente numerose. I quartili si ottengono suddividendo i dati ordinati in quattro parti ugualmente numerose.